DOCUMENT RESUME

ED 063 983 RC 006 165

AUTHOR Littlefield, John H.
TITLE The Use of Norm-Referenced Survey Achievement Tests with Mexican-American Migrant Students: A Literature Review and Analysis of Implications for Evaluation of the Texas Migrant Education Program.
SPONS AGENCY Texas Education Agency, Austin.
PUB DATE [72]
NOTE 51p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Achievement Tests; Elementary Grades; *Literature Reviews; *Mexican Americans; *Migrants; *Norm Referenced Tests; Reading Readiness; Surveys; Tables (Data)

ABSTRACT

The literature concerning the appropriateness of 9 norm-referenced survey achievement tests for use with Mexican American migrant students in grades 1 through 7 in Texas is reviewed in this report, which provides an evaluation of each test by the Center for the Study of Evaluation. The report provides the following information for each test: (1) ratings in the areas of math, reading, and oral-aural language; (2) the National Consortia for Bilingual/Bicultural Education report on the number of Title VII (Elementary and Secondary Education Act) bilingual projects which used the test during 1970-71; and (3) a review of research studies and state departments of education reports related to using the test with Mexican American or migrant students. Part II of the report discusses some of the implications of using norm-referenced tests and suggests possible alternative solutions to the problem of finding an appropriate instrument for evaluating Texas migrant education programs. Also included is a 42-item bibliography. (Author/NQ)

ED 063983

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

# The Use of Norm-Referenced Survey Achievement Tests with Mexican-American Migrant Students: A Literature Review and Analysis of Implications for Evaluation of the Texas Migrant Education Program

Prepared for the Texas Education Agency

by

John H. Littlefield

[1972]

RC 006165

## Abstract

This paper reviews the literature concerning the appropriateness of nine norm-referenced survey achievement tests for use with Mexican-American migrant students. An evaluation of each test by the Center for the Study of Evaluation is also included. Part II of the paper examines the implications of using norm-referenced tests to evaluate the Texas Migrant Education Program.

# INTRODUCTION

The following review represents an extensive effort to locate all available research studies comparing or evaluating the appropriateness of various norm-referenced achievement tests for measuring migrant students in grades 1 through 7. Since 98 percent of migrant students in Texas (TEA, 1971) are Mexican-American, the scope of the review was broadened to include studies of Mexican-American performance on these tests. While it is apparent that definite cultural differences exist between migrant and "typical" nonmigrant Mexican-American children, the generalization to all Mexican-American children is necessitated by the dearth of research material concerning the migrant child's performance on norm-referenced tests.

The following sources of research information were utilized:

1) ERIC (Educational Resources Information Center) was searched by the computer using the following descriptors: Standardized tests, norm-referenced tests, culture free tests, achievement tests, measurement instruments, testing programs or test reviews cross referenced with migrant children, migrant child education, migrant schools, Mexican-Americans or educationally disadvantaged.

2) The Education Index, a cumulative subject index to over 250 selected educational periodicals, proceedings and yearbooks, was searched from July 1965 to March 1972.

3) The U.S. Superintendent of Documents Monthly Catalogs, an index of publications by the U.S. Government Printing Office, was reviewed from January 1968 to March 1972.

4) CIJE (Current Index to Journals in Education), a monthly index of over 500 educational periodicals, was searched from its first issue in January 1969 through February 1972.

5) State Departments of Education in Arizona, California, Colorado, Florida, Nevada, and New Mexico were asked to recommend tests for use with Mexican-American migrant students.

6) The National Consortia for Bilingual/Bicultural Education, the Southwest Educational Development Laboratory, and numerous individuals who are working in the area of measuring Mexican-American and migrant educational achievement were asked to recommend tests.

Part I of the report will review nine major achievement tests which were recommended by the above sources or which have been included in the Anchor Test Study (ETS, 1972). For each test the following information will be provided:

(1) An evaluation of the test in the area of math, reading and oral-aural language from the CSE Elementary School Test Evaluations (Hoepfner, 1970).

(2) A report from the National Consortia for Bilingual/Bicultural Education (NCBE) concerning the number of ESEA Title VII Bilingual Projects which used the

test during the 1970-71 school year. (NCBE, 1971)

(3) A review of research studies which have used the test with Mexican-American or migrant students including reports from state departments of education which have used the test for measuring Mexican-American migrant students.

In order to select a norm-referenced evaluation instrument for migrant education programs three essential requirements must be met. (1) The test must be a well made measurement instrument as determined by standard psychometric criteria. The CSE Elementary School Test Evaluations (Hoepfner, 1970) have evaluated the nine tests to be reviewed on four basic criteria and thus will provide a comparative measure of the various instruments. (2) Second, the test must be appropriate for Mexican-American migrant students. This is a complex question currently being debated in educational journals. To provide a measure of the appropriateness of each test two criteria will be utilized: (1) A report of the number of ESEA Title VII Bilingual Projects which use the test as well as use by state departments of education with Mexican-American or migrant students; and (2) a review of research studies from the sources noted earlier that have used the tests with Mexican-American or migrant students. While a consensus judgment from the "authorities" is impossible at present, the above criteria should provide some insight into

the appropriateness of each test for use in measuring Mexican-American migrant students.

(3) The third and perhaps foremost requirement for selecting an evaluation instrument is that the tests measure the curriculum areas in which cognitive changes in the student are sought. The Texas Migrant Education Program is seeking changes in the areas of math, reading, and oral language. Math and reading are basic instructional areas measured by every achievement battery reviewed here. Oral language is a complex educational objective which has not been clearly defined in the Texas Migrant Education Program. At a very basic level it consists of oral and aural skills. The CSE Elementary School Test Evaluations (Hoepfner, 1970) have evaluated the tests reviewed here on oral-aural skills as an objective in measuring the educational goal called "reading" (Hoepfner, 1970, p. xii). In evaluating over 175 elementary school tests CSE did not classify a single test as measuring the oral skill entitled "speaking." Three tests reviewed here have been classified as measuring the aural skill called "listening-reaction and response." These evaluations will be noted in the discussions of individual tests, however at the outset it is important to recognize the difficulty of evaluating cognitive changes in an area broadly defined as oral language.

The content validity of a test for evaluation of a migrant

educational program must be determined by comparing test objectives and content to the objectives and curriculum of the program to be evaluated. Due to the wide variations of educational objectives and curriculum in Texas migrant education programs, no attempt will be made to assess the content validity of the tests reviewed in regard to specific program objectives and curricula. This can be done only at the individual project level or else await the adoption of a standard set of educational objectives and curriculum at the statewide level.

Part II of the report will discuss some of the implications of using norm-referenced tests for evaluation and suggest possible alternative solutions to the problem of finding an appropriate instrument for evaluating Texas migrant education programs.

PART I

Before beginning the test reviews a description of the test evaluations by the Center for the Study of Evaluation (CSE) is necessary. The CSE Elementary School Test Evaluations have evaluated "all the output measures that are prepared for or are potentially useful for evaluations within the elementary school, and that are generally available to educators and researchers" (Hoepfner, 1970, p. ix) on four assessment criteria:

(1) Measurement Validity--Each test was classified by CSE as to its educational goals then evaluations on the criterion of measurement validity were made in answer to the question: "Does the test appear to measure the specific educational objectives?"

(2) Examinee Appropriateness--This criterion assessed the appropriateness of the test's level of comprehension, its physical format, and the required response mode. Considerations such as quality of illustrations, size of print and spacing of materials on a printed page were included in assessing this criterion.

(3) Administrative Usability--This criterion addressed test utilization questions such as the practicality of group administration, simple and objective scoring procedures, and most important, ease of interpreting

test scores correctly. Finally the breadth and representativeness of the normative sample was evaluated.

(4) Normed Technical Excellence--The reliability, replicability, and refinement of measurement were the concern of this last criterion.

For each test objective selected by CSE, specific tests or subtests within a battery were evaluated using a 0-15 point scale for each of the four criteria discussed above. (E.g.: If the Reading Test of the Metropolitan Achievement Test (MAT) was classified as measuring "Total Reading," "Word Discrimination," and "Word Knowledge" then the CSE evaluations would evaluate the MAT Reading Test concerning how well it assessed student achievement on each of the three objectives. Each objective receives up to 15 points on each of the four criteria.) A grade of "good" corresponds to a scale value of 12-15 points on a particular criterion. "Fair" corresponds to 8-11 points and "Poor" corresponds to 0-7 points. A typical score for an objective might be F, F, P, B. CSE indicates that a "poor" rating renders a test clearly unsatisfactory on that particular criterion and therefore a better device should be sought to measure the educational objective being evaluated.

By utilizing this single evaluation source one gains three advantages: (1) Objectivity--CSE has no relationships with test publishers; (2) Conciseness--through the use of numerical scales

the various tests are easily compared; and (3) Consistency--it is rare in educational research to find a single set of criteria against which so many tests have been compared. Taken together, the four criteria used by the CSE evaluations provide an overall rating of how well these achievement tests assess educational objectives in the areas of reading and math and to a lesser extent the development of oral language.

## Test Reviews

(1) The Metropolitan Achievement Tests (MAT), 1958 edition, are a battery of language arts and mathematics tests based on traditional curriculum ranging from kindergarten to the ninth grade. The Metropolitan Readiness Test (MRT) preschool student readiness to learn reading and numbers. Table 1 outlines the ratings given to the Metropolitan tests designed for grades 1, 3, and 5 in the areas of math, reading, and aural language. The CSE 6th grade evaluations have been omitted since the MAT intermediate level tests are designed for both the 5th and 6th grades.

Table 1

Metropolitan Readiness and Achievement Tests[a]

| Objective of the Test | Measurement Validity | Examinee Appropriateness | Administrative Usability | Normed Technical Excellence |
|---|---|---|---|---|
| 1st Grade | | | | |
| Readiness Test: | | | | |
| Spatial Reasoning | F | F | F | F |
| Comprehension of Numbers | F | G | G | P |
| Listening Reaction & Response | G | F | F | P |
| Phonetic Recognition | F | F | F | P |
| Recognition of Word Meaning | G | F | F | P |
| Arithmetic--Concepts | F | F | F | F |
| Reading: | | | | |
| Word Discrimination | F | F | F | F |
| Word Knowledge | F | F | F | F |
| Total | P | F | F | F |
| 3rd Grade | | | | |
| Arithmetic: | | | | |
| Computation | P | F | G | F |
| Problem Solving & Concepts | P | F | G | P |
| Reading: | | | | |
| Comprehension | F | F | G | F |
| Word Discrimination | F | F | G | F |
| Word Knowledge | F | F | G | F |
| 5th Grade | | | | |
| Arithmetic: | | | | |
| Computation | F | F | G | F |
| Problem Solving & Concepts | F | F | G | F |
| Reading: | | | | |
| Comprehension | F | F | G | F |
| Word Knowledge | F | F | G | F |

[a]Information obtained from Hoepfner (1970).

Table 1 shows the MRT and MAT to be generally well-made tests. The MRT listening test is classified by CSE as an aural language measure. NCBE (1971) points out that MAT is used in nine bilingual education projects. The MRT on the other hand is utilized by sixteen bilingual projects.

Horn (1966) notes that the large number of zero scores attained by Spanish-speaking students on the MRT indicates that it is inappropriate for pretesting Spanish-speaking preschool children. Robison (1966) states that the MRT is reliable with culturally disadvantaged children (r = .907), however the weakness of the test lies in the complexities involved in test administration and the length of time needed for each administration. Mitchell (1967) indicated that the MRT's predictive validity, based on correlations with the Stanford Achievement Test, is virtually identical for Caucasian, Negro-, Oriental-, or Mexican-Americans. Davis and Personke (1968) found that Spanish-speaking school entrants performed equally well on the English or colloquial Spanish versions of the MRT and thus provided "no evidence on which to question the practice of administering tests in English to Spanish-speaking school entrants." In a later study of the same students, Personke and Davis (1969) found that the Alphabet subtest of the MRT administered in English was the best predictor of school achievement of Spanish-speaking students.

Turning to the Metropolitan Achievement Tests (MAT), Arnold (1969) administered the MAT level 1, designed for second grade students to disadvantaged bilingual students beginning the third grade. The test was found to be quite reliable (Alpha Coefficient = .95) with disadvantaged "bilingual" students who had two years of oral language instruction. Bordie (1969) in reviewing both the MAT and the MRT with regard to linguistically different learners stated that language material was inadequate for students at the lower end of the scales since chance scores are less than 1/2 grade below the minimum level for which norms are offered. Eagle and Harris (1969) studied the interaction effects between tests and socio-cultural variables. Although they contrasted white students with nonwhites (primarily Negroes), Eagle and Harris demonstrated that the Metropolitan Achievement Battery produced a greater discrepancy between races at the 4th and 6th grade levels than did the Iowa Test of Basic Skills. This interaction effect between tests and sociocultural variables is important in selecting an achievement measure for Mexican-American migrant students, however no other studies of this type have been found in the educational literature.

Hurt and Mishra (1970) found the Metropolitan Achievement Test to be as reliable for disadvantaged Mexican-American children as for the normative sample. Validity of the MAT was determined by using the Wide Range Achievement Test as the criterion

of validity. In discussing the findings Hurt and Mishra state, "MAT has relatively high validity for the sample groups." Johnson's (1971) analysis of the Metropolitan Tests revealed that they contain symbols not equally familiar and motivating to all socioeconomic groups. He points out that they are fixed on verbal symbols, paragraph content, and problem solving most familiar to middle and upper socioeconomic group children. Solomon (1971) found that response in a test booklet, response on a separate non-machine scorable form, and response on separate machine scorable forms had no effect on the scores of culturally deprived fourth grade students taking the MAT. Hutchison (1972) concluded that the Word Discrimination Test of the MAT uses "dialect-prejudiced items" and thus is not appropriate for children in urban ghetto areas.

A 1970 revision of the MAT has been published. Buros (1972) does not review the 1970 version, and refers readers to reviews of the 1958 edition. Thus the information presented here is apparently the most current evaluative material of the MAT available. The inclusion of only 4 percent Mexican-Americans, Cubans, and Puerto Ricans in the 1970 edition norming sample (Harcourt et al., 1971) indicates that the orientation of the new battery has not changed substantially since the 1958 edition which was also normed primarily on white middle class children (Durost, 1959).

(2) The Stanford Achievement Tests, 1964 revision, measure

achievement in reading, language arts, and arithmetic. The CSE evaluations of this test in the areas of arithmetic and reading are presented in Table 2.

Table 2

Stanford Achievement Test[a]

| Objective of Test | Measurement Validity | Examinee Appropriateness | Administrative Usability | Normed Technical Excellence |
|---|---|---|---|---|
| 1st Grade | | | | |
| Spelling | F | G | G | F |
| Arithmetic Total | P | F | G | F |
| Reading--Paragraph Meaning | P | F | G | P |
| " Word Meaning | F | G | G | P |
| " Vocabulary | F | F | G | P |
| 3rd Grade | | | | |
| Arithmetic--Concepts | F | F | G | F |
| " Computation | F | G | G | F |
| Reading--Paragraph Meaning | F | F | G | F |
| " Word Meaning | P | F | G | P |
| 5th Grade | | | | |
| Arithmetic--Concepts | F | F | G | F |
| " Computation | F | F | G | F |
| " Applications | F | F | G | F |
| Reading--Paragraph Meaning | F | F | G | F |
| " Word Meaning | F | G | G | P |
| 6th Grade | | | | |
| Arithmetic--Concepts | F | F | G | P |
| " Computation | F | F | G | F |
| " Applications | F | F | G | F |
| Reading--Paragraph Meaning | F | F | G | F |
| " Word Meaning | F | F | G | F |

[a] Information obtained from Hoepfner (1970).

Like the Metropolitan tests, the Stanford Achievement Tests are a well-built battery, rated particularly high in Administrative Usability but lower in Measurement Validity and Normed Technical Excellence. It also does not provide any measure of aural language. The test is used in thirteen bilingual projects (NCBE, 1971). Very little research work has been designed to evaluate the appropriateness of the Stanford Achievement Tests for use with Mexican-American or migrant children. Horn (1966) points out a major limitation of his study of reading-teaching methods for disadvantaged Spanish speaking 1st graders was the standardized instruments, namely the Metropolitan Readiness Test and the Stanford Achievement Test. Palomares and Cummins (1967) chose the Stanford series "after careful selection" to evaluate rural M-A pupils in preschool and grades 1-6. No further comment was made on the efficacy of the test with Mexican-American students. Hawkridge (1969) points out that the Malabar reading program for Mexican-American children found the floor of the Stanford Achievement Test: Reading to be too high for Mexican-American students. Bordie (1969) points out that the language portion of the Stanford is concerned with the measurement of ability in formal written English, an ability which is usually lacking in the linguistically different student.

The Texas Education Agency has used the Stanford Achievement Test since 1965 to evaluate migrant education programs.

Buros (1972) reports that the Stanford Early School Achievement Test (SESAT) published in 1969-70 measures grades K through 1.5 with four subtests: Environment, Mathematics, Letters and Sounds, and Aural Comprehension. Buros points out the test is not a "readiness test." No research material concerning the appropriateness of the SESAT for use with Mexican-American or migrant students was located.

(3) California Achievement Tests (CAT), 1957 edition with 1963 norms, consist of tests of knowledge and understanding in Reading, Arithmetic, and Language. There are alternate forms for each level. The CSE evaluation of this test in the areas of math and reading are outlined in Table 3.

Table 3

California Arithmetic Test and California Reading Test[a]

| Objective of the Test | Measurement Validity | Examinee Appropriateness | Administrative Usability | Normed Technical Excellence |
|---|---|---|---|---|
| **1st Grade** | | | | |
| Arithmetic Reasoning--Total | P | F | F | P |
| Arithmetic Fundamentals--Total | F | F | F | F |
| Vocabulary--Total | F | F | G | P |
| " Word Recognition | P | F | G | P |
| " Word Form | F | F | F | P |
| " Meaning of Opposites | P | F | F | P |
| " Picture Association | P | F | F | P |
| Reading Comprehension | P | F | F | P |
| **3rd Grade** | | | | |
| Arithmetic Reasoning--Total | F | F | F | F |
| Arithmetic Fundamentals--Total | G | F | G | F |
| Reading Comprehension--Total | F | F | G | F |
| Reading Vocabulary--Total | P | F | G | F |
| Reading Vocabulary: | | | | |
| Meaning of Opposites | F | F | F | P |
| Word Recognition | P | F | G | F |
| **5th Grade** | | | | |
| Arithmetic Reasoning--Total | P | F | G | P |
| Mathematics Vocabulary | P | F | F | P |
| Arithmetic Fundamentals--Total | G | F | F | F |
| Arithmetic Reasoning--Total | P | F | G | P |
| General Vocabulary | P | F | F | P |
| Reading Vocabulary | F | F | F | F |
| Reading Comprehension--Total | P | F | G | F |
| **6th Grade** | | | | |
| Arithmetic Reasoning--Total | P | F | G | P |
| Mathematics Vocabulary | P | F | F | P |
| Arithmetic Fundamentals--Total | G | F | F | F |
| Reading Comprehension | P | F | G | F |
| General Vocabulary | P | F | F | F |
| Reading Vocabulary | F | F | F | F |

[a]Information obtained from Hoepfner (1970).

A survey of Table 3 indicates that the CAT was rated somewhat lower than the Metropolitan and Stanford Achievement Tests. Measurement Validity is particularly low in comparison to other tests reviewed here. No measures of aural language have been included.

NCBE (1971) indicates that the CAT is used in seven bilingual projects. Very few research studies of the appropriateness of the CAT for measuring Mexican-American or migrant students were located. Two studies (Perrodin & Snipes, 1966; and Morris, Pestaner, and Nelson, 1967) used the CAT to measure the effects of mobility on student achievement, however neither study commented on the adequacy of the test for this purpose. Atilano (1971) used the CAT to evaluate the 1970-71 Grants, New Mexico Bilingual Project. Atilano indicated that the CAT provided "noteworthy results" in measuring achievement gains by Mexican-American, Indian and white elementary school children. This type of information is not a satisfactory substitute for empirical studies of the appropriateness of the CAT for measuring Mexican-American students, however it does provide some information concerning the utility of the CAT in practical evaluation applications.

In 1970 a revision of the CAT was published. Green (1972) described the 1970 edition as "basically similar to the 1957 edition" and found that Mexican-Americans in the southwestern

United States had a high overall median reliability (KR 20 = .90) on the CAT - 70. Using a unique approach to test bias as caused by item selection methods utilized in constructing norm-referenced achievement tests, Green studied data collected during the standardization of the CAT, 1970 edition. He concluded that selecting test items based on tryouts of populations which are primarily white and middle-class results in the selection of items which are more biased against minority groups than against white, middle-class children. This phenomenon was held to be true for most published batteries of standardized tests.

Buros (1972) does not review the CAT - 70 and refers readers to reviews of the 1957 edition. No further evaluative information is available at this time.

(4) The Comprehensive Tests of Basic Skills (CTBS), 1968 edition, measure Reading, Language, Arithmetic and Study Skills at grades 2.5 - 12. The test is available in two forms at four overlapping levels with similar content at each level. The CSE evaluations of the test for the areas of arithmetic and reading are presented in Table 4.

Table 4

Comprehensive Tests of Basic Skills[a]

| Objective of Test | Measurement Validity | Examinee Appropriateness | Administrative Usability | Normed Technical Excellence |
|---|---|---|---|---|
| *3rd Grade* | | | | |
| Arithmetic--Total | F | F | G | G |
| " Concepts | F | F | G | P |
| " Computation | F | G | G | F |
| " Applications | F | G | G | F |
| Reading--Comprehension | F | F | G | G |
| " Total | F | F | G | G |
| *5th Grade* | | | | |
| Arithmetic--Concepts | P | F | G | F |
| " Computation | F | G | G | F |
| " Total | F | G | G | G |
| " Applications | F | F | G | P |
| Reading--Comprehension | F | F | G | G |
| " Total | F | F | G | G |
| " Vocabulary | F | F | G | F |
| *6th Grade* | | | | |
| Arithmetic: | | | | |
| Concepts | P | F | G | F |
| Total (Form 2Q) | F | G | G | G |
| Total (Form 3Q) | F | F | G | F |
| Applications (Form 2Q) | F | F | G | P |
| Applications (Form 3Q) | F | F | G | P |
| Reading--Total | F | F | G | F |
| " Vocabulary | P | F | G | F |
| " Comprehension | F | F | G | P |

[a]Information obtained from Hoepfner (1970).

The CTBS in general have received a high rating particularly in Administrative Usability. No measure of aural language has been included.

NCBE (1971) indicates that the CTBS is used in one Title VII Bilingual Project. No research studies concerning the appropriateness of the CTBS for measuring Mexican-American or migrant students were located, however since the test was first published in 1968 its newness may account somewhat for its relative lack of use. The California Department of Education reported (California, 1972) that it used the CTBS to measure migrant students in grades 4 - 8.

(5) Iowa Test of Basic Skills (ITOBS), 1955-56 edition, are tests of generalized intellectual skills in five major areas: vocabulary, reading, language, work-study, and arithmetic. The vocabulary, reading, and arithmetic evaluations by CSE make up Table 5.

Table 5

Iowa Tests of Basic Skills[a]

| Objective of Test | Measurement Validity | Examinee Appropriateness | Administrative Usability | Normed Technical Excellence |
|---|---|---|---|---|
| 3rd Grade | | | | |
| Arithmetic Skills: | | | | |
| Total | F | F | G | G |
| Arithmetic Concepts | F | F | G | G |
| Problem Solving | F | F | G | F |
| Reading--Comprehension | F | F | G | F |
| Vocabulary | F | F | G | G |
| 5th Grade | | | | |
| Arithmetic Skills: | | | | |
| Total | F | F | G | F |
| Arithmetic Concepts | F | F | G | F |
| Problem Solving | F | F | G | F |
| Reading--Comprehension | F | F | G | G |
| Vocabulary | F | F | G | G |
| 6th Grade | | | | |
| Arithmetic Skills: | | | | |
| Total | F | F | G | G |
| Arithmetic Concepts | F | F | G | F |
| Problem Solving | F | F | G | F |
| Reading--Comprehension | F | F | G | G |
| Vocabulary | F | F | G | G |

[a]Information obtained from Hoepfner (1970).

Table 5 shows the CSE evaluation of the ITOBS to be very high. The absence of any "Poor" ratings was not found in any other CSE evaluation reviewed here. Again no aural language measure has been provided.

NCBE (1971) points out that only two Title VII Bilingual

projects were using the ITOBS in 1970. Only one relevant research study was available (Eagle and Harris, 1969) pertaining to the appropriateness of using the ITOBS with culturally disadvantaged students. This study was cited earlier (page 12) concerning the relative advantage of Negro students (differences between mean scores of blacks and whites were smaller) on the ITOBS as compared to the Metropolitan Achievement Tests. Buros (1965, p. 13) indicates that the ITOBS has been widely used in comparative studies of achievement tests thus its lack of use with Mexican-Americans or migrants is difficult to explain.

A 1970 edition of ITOBS is now available, however it is not reviewed in Buros (1972) and no evaluative information comparable to the CSE evaluations is available at this time so inferences concerning its qualities can only be made from evaluations of the 1956 edition.

(6) Sequential Tests of Educational Progress (STEP), 1956-57 edition, is a battery of general achievement tests in six academic areas: Reading, Writing, Listening, Mathematics, Science and Social Studies. The CSE evaluations for the STEP are presented in Table 6.

Table 6

Sequential Test of Educational Progress[a]

| Objective of Test | Measurement Validity | Examinee Appropriateness | Administrative Usability | Normed Technical Excellence |
|---|---|---|---|---|
| 5th Grade | | | | |
| Mathematics | F | F | F | P |
| Listening: Reaction & Response | F | F | F | F |
| Reading | F | F | F | F |
| 6th Grade | | | | |
| Mathematics | F | G | F | P |
| Listening: Reaction & Response | F | G | F | P |
| Reading | F | F | F | P |

[a]Information obtained from Hoepfner (1970).

The STEP did not receive particularly high evaluations. It does not provide measures below grade 4, however it uses a format similar to the Cooperative Primary Tests (grades 1-3) which will be reviewed later. The tests of listening provide a measure of aural language related to the educational objective entitled "listening--reaction and response."

NCBE (1971) points out that the STEP is used in two Title VII Bilingual Projects. No research studies were found concerning the appropriateness of the STEP for measuring Mexican-American or migrant students. Bordie (1969) in a review of language tests for linguistically different learners states that the test "discriminates against students at both ends of the

ability range." A 1969 edition entitled STEP--Series II has been published, however it does not include the listening test offered in the original series. Buros (1972) does not review the STEP--Series II thus the CSE evaluation remains the most current evaluative material on the STEP.

(7) SRA Achievement Series, 1954-64 edition, measures mastery of content in the areas of Reading, Arithmetic, Language Arts, Social Studies, Science and Work Study Skills. The CSE evaluation of the SRA series in Reading and Arithmetic are outlined in Table 7.

Table 7

SRA Achievement Series[a]

| Objective of Test | Measurement Validity | Examinee Appropriateness | Administrative Usability | Normed Technical Excellence |
|---|---|---|---|---|
| 1st Grade | | | | |
| Arithmetic--Concepts | F | G | G | F |
| " Computation | P | G | G | F |
| " Total | P | F | F | F |
| " Reasoning | F | G | G | P |
| Reading: | | | | |
| Language Perception | F | F | G | F |
| Comprehension--Form C | F | F | G | P |
| Comprehension--Form D | P | F | F | P |
| Total | P | F | F | F |
| Verbal Pictorial Comprehension | P | G | G | F |
| Vocabulary--Form C | F | G | G | P |
| Vocabulary--Form D | P | G | G | F |
| 3rd Grade | | | | |
| Arithmetic--Concepts | F | G | G | P |
| " Computation | F | G | G | F |
| " Reasoning | F | G | G | P |
| Reading--Comprehension | F | F | G | P |
| " Total | F | G | G | F |
| " Vocabulary | F | F | G | F |
| 5th Grade | | | | |
| Arithmetic--Concepts | F | F | G | G |
| " Computation | F | F | G | F |
| " Reasoning | F | F | G | F |
| Reading--Total | F | F | G | F |
| " Vocabulary | F | F | G | G |
| " Comprehension | F | F | G | F |
| 6th Grade | | | | |
| Arithmetic--Concepts | F | F | G | G |
| " Computation | F | F | G | F |
| " Reasoning | F | F | G | F |
| Reading--Total | F | F | G | F |
| " Vocabulary | F | F | G | G |
| " Comprehension | F | F | G | P |

[a]All information obtained from Hoepfner (1970).

The SRA series received a fairly high evaluation from CSE compared to other tests in this review. No measure of aural skills has been included.

NCBE (1971) indicates that the SRA Achievement Series is used in two Title VII Bilingual Education Projects. No research work was found concerning the appropriateness of the SRA Achievement Series for measuring Mexican-American or migrant students. A 1970 edition of the SRA Achievement Series has been published. Buros (1972) describes the new Forms C and D to be "of generally high quality" and points out that corrections and refinements of the earlier forms make these forms "better than their predecessors."

(8) The Inter-American Series include two separate tests, the Tests of General Ability (TOGA) and the Tests of Reading (TOR). Unlike most of the tests reviewed here, the Inter-American Series is not an achievement battery. The TOGA provides an estimate of ability to do academic work, but is too limited in breadth, according to the author, to provide an adequate measure of general intelligence (Manuel, 1967, p. 4). The TOR is used as a measure of reading achievement and also as a basis for estimating ability to do school work in other areas in which the ability to read is related to achievement (Manuel, 1967, p. 4). Both tests are available in equivalent English and Spanish versions. The CSE evaluations of the English language editions of the Inter-American Series is outlined in Table 8.

Table 8

Inter-American Series[a]

| Objective of Test | Measurement Validity | Examinee Appropriateness | Administrative Usability | Normed Technical Excellence |
|---|---|---|---|---|
| **Test of General Ability** | | | | |
| 1st Grade | | | | |
| Verbal--Numerical | P | F | F | P |
| 3rd Grade | | | | |
| Total | F | F | G | F |
| 5th Grade | | | | |
| Numerical--Total | F | F | F | F |
| Verbal | F | F | F | P |
| 6th Grade | | | | |
| Numerical--Total | F | F | F | F |
| Verbal | F | F | F | P |
| **Test of Reading** | | | | |
| 1st Grade | | | | |
| Total | P | F | F | P |
| Vocabulary | F | F | F | P |
| Comprehension | P | F | F | P |
| 3rd Grade | | | | |
| Speed | F | F | F | P |
| Total | F | F | F | F |
| Vocabulary | F | F | F | P |
| Comprehension | F | F | F | P |
| 5th Grade | | | | |
| Speed | F | F | F | P |
| Total | F | F | F | F |
| Vocabulary | F | F | F | P |
| Level of Comprehension | F | P | F | P |
| 6th Grade | | | | |
| Speed | F | F | F | P |
| Total | F | F | F | F |
| Vocabulary | F | F | F | P |
| Level of Comprehension | F | P | F | P |

[a]Information obtained from Hoepfner (1970).

The TOGA and TOR did not receive particularly high evaluations relative to other tests reviewed. No measures of aural skills have been included, however a separate test, Comprehension of Oral Language Test in Spanish, is available.

NCBE (1971) points out that six Title VII Bilingual Projects use the Spanish edition of the TOGA, entitled Prueba de Habilidad General, while twelve subjects use the English edition of the TOGA. The Spanish edition of the TOR, entitled Prueba de Lectura, is used by fourteen bilingual projects while the English edition is used by five projects. Manuel (1967, p. 11) administered the English and Spanish versions of the TOGA to Spanish-speaking 2nd grade students in El Paso. Median total scores on the two editions differed by only .4 of a point (53.7 - 53.3), however the Spanish version produced a wider range of scores (28 - 82) than the English version (33 - 72). Arnold (1969) showed that the English version of the TOR was more reliable in terms of internal consistency (Coefficient Alpha = .91 for total test) than was the Spanish version (Coefficient Alpha = .72 for total test) when used with 3rd grade disadvantaged bilingual students. Manuel (1967, p. 14) recommends that the Inter-American tests be used with local or regional norms. For Spanish-speaking students in English-language schools he recommends that the student be tested in the language he knows best or perhaps in both languages. Manuel provides tentative norms for students of New York City

schools in which English-speaking students took the English edition while Spanish-speaking students took the Spanish edition (Manuel, 1967, p. 39). In a pilot attempt to apply evaluation procedures in the classroom, Melarago and Newmark (1968) showed that after a year of formal reading instruction white students (N = 26) obtained a mean score of 52.4 on the Gates Primary Reading Test while Mexican-Americans (N = 25) scored a mean of 13.0. Despite the wide difference in measured reading ability between the two groups the TOGA showed no significant score differences on the Oral Vocabulary and Numbers subtests. This finding supports Manuel's claim discussed earlier that the test measures the ability to do academic work, but is not necessarily a measure of achievement.

In an evaluation of a Bilingual Education Program, Bortin (1970) notes that on several items of the TOGA, disadvantaged bilingual students in Milwaukee misinterpreted the pictures with which a written word was to be associated. The children were confused by the small size and detail of the pictures and the large number of pictures per page. The TOR levels 1 and 2 was criticized for being too long and difficult for the achievement level of migrant students in grades 1 - 3. Evaluator comments such as these provide an invaluable source of information concerning the appropriateness of various batteries for use with Mexican-American or migrant students. A review of evaluation

reports of migrant programs listed in the ERIC files revealed that while most evaluators were not satisfied with the measurement instruments used, no specific criticisms were made.

(9) Cooperative Primary Tests (CPT) are designed to measure basic verbal and quantitative understandings in grades 1 - 3. As noted earlier, the CPT is designed to be used in conjunction with the Sequential Tests of Educational Progress, Series II, to measure students from grades 1 - 14. The battery consists of six tests: a Pilot test for grades 1 and 2 to provide the children with practice in handling tests materials; a Listening test in which the child marks pictures in his test booklet in response to words, sentences, stories, and poems read by the teacher; Word Analysis; Mathematics; Reading; and Writing Skills. The CSE evaluation of the CPT in the areas of Mathematics, Listening, Word Analysis, and Reading are outlined in Table 9.

Table 9

Cooperative Primary Test[a]

| Objective of Test | Measure-ment Validity | Examinee Appropri-ateness | Adminis-trative Usability | Normed Technical Excellence |
|---|---|---|---|---|
| 1st Grade | | | | |
| Mathematics | F | P | G | F |
| Listening | F | F | G | F |
| Word Analysis: Phonetic Recognition | F | F | G | F |
| Reading | P | F | G | F |
| 3rd Grade | | | | |
| Mathematics | P | F | G | F |
| Listening: Reaction & Response | F | G | G | F |
| Reading | P | F | G | F |

[a]Information obtained from Hoepfner (1970).

Table 9 indicates that CSE evaluated the CPT fairly highly except on the criterion of measurement validity. The CPT listening test is classified as a measure of aural language.

NCBE (1971) reports that the CPT battery of individual tests, namely Reading and Mathematics, are used in nine Title VII Bilingual Education Projects. The California Department of Education uses the CPT for measuring migrant students in grades 2 and 3 (California, 1972). No research work concerning the appropriateness of the CPT for measuring Mexican-American or migrant students was found. Pickering (1969) used the test in a study of intellectual abilities of culturally disadvantaged first-grade

children as predictors of achievement in reading, mathematics and listening. Pickering's complete text was unavailable at this writing thus no evaluative statements on the CPT were available.

## Summary

The nine tests reviewed here provide a wide cross section of standardized tests used by school systems throughout the United States. Data collected by the U.S. Office of Education on test usage indicate that the seven tests of the Anchor Test Study (tests 1- 7 reviewed here) are used with more than 90 percent of the 4th, 5th, and 6th grade children tested by school systems in the United States (ETS, 1971). Obviously, many other tests have been used in measuring Mexican-American or migrant students, however a review of the commercial tests used in Title VII Bilingual projects indicates that all of the widely used achievement measures (five or more projects) have been included in this review.

Several authors (De Avila, 1972; Bernal, 1972) have called for a Buros-type review of tests for use with Mexican-American children. Correspondence with both these authors indicates that no such reviews are forthcoming in the near future. At the present time no other authors appear to have attempted to collect the available information concerning the appropriateness of

these standardized norm-referenced measures for use with Mexican-American or migrant students.

## PART II

Upon completing a review of the literature on survey achievement tests used with Mexican-American migrants, one is prompted to question whether current practices are the most relevant available methods of evaluating educational programs. Standardized achievement tests are often said to be biased against and thus inappropriate for children belonging to disadvantaged racial and ethnic minority groups (Williams, 1970; Houston, 1971). Green (1972) points out that research on bias in achievement tests is essentially nonexistent. The limited number of studies reviewed in Part I testifies to a neglect of this area by educational researchers. However, before condemning or condoning current achievement tests an assessment of alternative methods of educational evaluation is in order.

Criterion-referenced testing is a popular approach now supported by many educational theorists. It offers the advantages of high content validity and relative freedom from charges of test bias since all test items in the ideal situation would have been specifically taught in the classroom. In addition, students are compared against a specified criterion or against their own previous performances rather than competing with one another. Cronbach (1970) points out that the testing movement has given too much attention to comparative judgments and too

little to absolute, content-referenced measurement. Evaluation of educational programs calls for absolute judgments concerning whether or not participants exhibit the desired behaviors after experiencing instruction. Thus criterion-referenced tests are a more adequate approach to evaluating educational programs than are norm-referenced measures. Popham and Husek (1969) point out that for the evaluation of treatments "norm-referenced measures are not the most suitable devices for such purposes since their emphasis on producing heterogeneous performance sometimes diverts them from adequately reflecting the treatment's intended objectives." Norm-referenced tests are designed for selection situations in which one is interested in determining how students performed in comparison to a norming group.

At this point we must carefully consider our objectives in selecting a test to evaluate migrant education programs. Initially we want to know whether students in a particular program are achieving the educational objectives of the statewide migrant education program. However, within the larger context of the goals of a remedial educational program we are ultimately interested in preparing the migrant child to adequately take his place in standard public school classrooms. Even though a student may have mastered the educational objectives of the statewide or local migrant education program, the program has not been successful until he performs satisfactorily in classrooms

aimed at the dominant cultures of our society. To propose that special classes for migrant children will be necessary throughout their education is equivalent to endorsing "separate but equal education" and is not likely to receive wide acceptance by either the dominant or minority ethnic and cultural groups of the United States. Thus the selective nature of norm-referenced tests can provide useful information about a migrant student's potential for achievement within a broader educational setting than the migrant program. As individual students become able, it is imperative that they be allowed to enter "regular" classrooms serving the dominant cultural and ethnic groups.

Returning to a consideration of the various facets of criterion-referenced testing in its present state of development, several important limitations must be examined. First, the exact definition of the term criterion-referenced test is still being debated. Glaser (1963) has been credited with introducing the term. He defined the criterion-referenced test performance as being a behavioral statement (or statements) that is made without reference to the performance of other individuals (e.g., the student can multiply 3-digit numbers together correctly). Wang (1969) described a criterion-referenced test as "an achievement test developed to assess the presence or absence of a specified criterion behavior described in an instructional objective." Jackson (1971) held both of the above definitions to be inadequate

and added that the term criterion-referenced test applied "only to a test designed and constructed in a manner that defines explicit rules linking patterns of test performance to behavioral referents." Ebel (1971) pointed out that percentage-mastery grades, widely favored in schools and colleges in the U.S. at the turn of the century, represent one form of criterion-referenced measurement. By using different interpretations of the term criterion-referenced test one embraces different testing implications which he may or may not be prepared to deal with.

Second, in devising a criterion-referenced achievement evaluation instrument one must make certain value judgments concerning which educational objectives are essential outcomes of the instruction provided. This value judgment is made in all achievement testing situations. Norm-referenced tests provide broad coverage of subject areas by making inferences about certain behavioral domains from performance on related tests (e.g., vocabulary provides one measure of reading). However, the direct task-sampling approach of criterion-referenced testing requires that test questions be selected directly from the domain of behaviors taught thus limiting the amount of content that can be reasonably covered by an achievement evaluation instrument. Popham and Husek (1969) point out that for evaluating the adequacy of treatments with criterion-referenced tests, different

people can take different test items thereby sampling a wide domain behavior with numerous short tests. In this situation however, as pointed out earlier, we are also interested in making judgments about which individuals are prepared to enter "regular" public school classrooms.

The validity of an evaluative instrument depends on the value of the question: "Did the test appraise the educational objectives I consider most important?" This question obviously will receive different answers from different people. Cronbach (1970) points out that "an ideally suitable battery for evaluation purposes will include separate measures of all outcomes the users of the information consider important." The important consideration here is that norm-referenced tests assess a broad range of behavioral traits and abilities. A criterion-referenced test on the other hand will necessarily neglect many areas which some decision makers consider crucial to evaluating the success of the education program in preparing a student to participate in "regular" classrooms.

A third important limitation of criterion-referenced tests is the inapplicability of traditional procedures for determining test reliability and validity plus differences in item analysis, score reporting, and score interpretation. Popham and Husek (1969) address these implications of criterion-referenced tests directly. They point out that the issue of variability of

scores is at the core of the difference between norm-referenced and criterion-referenced tests. Efforts are being made to develop statistical procedures and decision models for evaluating the test characteristics of criterion-referenced tests (Livingston, 1972; Edmonston et al., 1972); however the important consideration for this discussion is that while procedures are being developed, many areas have not been formalized to the extent that they are ready to be employed in a statewide educational evaluation program.

In summary, the foregoing arguments are not intended as an endorsement of norm-referenced testing. Many problems exist with our present methods of achievement assessment, however a large number of these problems stem from the uses which are made of norm-referenced achievement tests rather than from the information the tests provide. This is not a fault of the test. It is this author's contention that many people in education are endorsing criterion-referenced testing as a universal solution to the traditional problems of testing (e.g., bias against minority groups, inappropriateness to local educational objectives). It does hold promise for vastly improving and augmenting our present methods of measurement and evaluation. However at present, the methodology for utilizing and interpreting criterion-referenced tests has not been formalized or disseminated to public school teachers and educational field workers. These

limitations must be carefully considered before selecting criterion-referenced testing, on a statewide basis, to evaluate Texas migrant education programs.

At this point one is faced with the dilemma of accepting standardized norm-referenced survey achievement tests as described in Part I of this paper or delving into the area of criterion-referenced testing with its concomitant problems and limitations as described above. Fortunately an intermediate position exists.

Green (1972) indicates that bias in standardized achievement tests stems from two possible sources. (1) The thoughts and preconceptions of the item writers are different from those of cultural minority children who will take the tests. (2) The customary item tryout and selection procedures use groups who are different from minority children thus the items finally selected are biased and discriminate against groups unlike the modal group in the tryout sample. The bias introduced by the thoughts and preconceptions of item writers is difficult to deal with. Item writers might be trained to avoid using culturally-biased contexts, however that is a complex requirement which also entails writing new tests. Item tryout and selection procedures on the other hand can be changed by selecting items which correlate best with total test scores of minority group students. This is one procedure for making currently existent

norm-referenced survey achievement tests more responsive to relevant skills and abilities of Mexican-American students. Green (1972) studies this procedure in detail.

Using the data obtained during the standardization of the California Achievement Tests, 1970 Edition (CAT - 70), Green attempted to determine whether using seven different item tryout groups, including: residential/suburban whites in the north, central city blacks in the north, residential/suburban whites in the south, rural blacks in the south, small and large city Mexican-Americans in the southwest, and city and suburban whites in the southwest, would lead to selection of different test items. The original items were described as being "written by and for 'middle America.'" The basic procedure was to treat each group as a tryout sample with the items of the CAT - 70 serving as the item pool. For each group on each test of the battery at the 1st, 3rd, 5th, 8th, and 10th grade levels, the best half of the items (i.e., those with the highest item-test correlations) were noted. Four kinds of analyses were made, however only two analyses involving comparisons between Mexican-Americans and Anglo-Americans in the southwest will be presented here.

1) The number and percent of items chosen for one group but not the other. These items were labeled "biased" since the proportion of items falling in this group

indicates the degree to which the two groups interact in a distinct manner with the test items. The .43 median proportion of biased items between Mexican-Americans and whites in the southwest was the highest median proportion of biased items found in any group comparison.

2) The mean scores on the full tests and the biased item half tests were examined for changes in the relative status of the group as a result of item selection. This looks at test bias from the standpoint that scores of a minority group are unfairly low because the test does not adequately measure all the relevant abilities or knowledge on which the group in question happens to score well.

If the item pool contains items which measure relevant attributes, then selecting the items biased toward Mexican-Americans should raise the mean scores of Mexican-Americans more from the full test to the half test than selecting items biased toward whites raises the mean scores of whites. This was found to be the case in 25 of 36 tests compared (nine tests at each of grades 1, 3, 5, and 8; $p = .02$) thus supporting the hypothesis that the full test was biased against Mexican-Americans who had relevant abilities which were not indicated in the full scores. In a final table using these same biased item half tests, Green

shows that differences between Mexican-American and white mean scores are less on the half test biased toward Mexican Americans in 27 of the 36 tests ($p = .01$).

The implications of this study for selecting a test to measure Mexican-American migrants are obvious. Rather than attempt to develop a whole new criterion-referenced test with all of the attending limitations and precautions or adopt a commercially available norm-referenced test which has not been shown to be appropriate for Mexican-Americans; a more plausible route would be to choose the best available achievement battery based on information contained here and adapt each test to migrant students using the procedures similar to those described by Green (1972). State norms for the half length instrument would satisfy the requirements of Public-Law 89-10 for using "some objective standard or norm" (ESEA, 1965) and provide a comparative measure of student achievement yet the instrument would be responsive to cognitive changes in Mexican-American migrant students. The choice of a test which closely reflects the educational objectives of the Texas migrant education program will provide evaluative information concerning how well educational objectives are being met. Frenner (1972) presents several useful techniques using the criterion of "minimal competency" to gain criterion-referenced interpretations of survey achievement tests. More importantly, the state norms and selective nature

of an "adapted" norm-referenced test will provide useful information concerning whether individual migrant students are ready to enter "regular" classrooms to continue their education.

In summary, the rejection of criterion-referenced testing for evaluation of the Texas migrant education program has not been advocated. To fully evaluate remedial education programs such as the migrant program, both norm-referenced and criterion-referenced tests are needed. At present, however, the lack of statewide educational objectives and the relative newness of criterion-referenced testing methodology severely restricts the appropriateness of this alternative for statewide employment. Certainly the development and pilot testing of criterion-referenced tests in the migrant program should begin. As an interim step to eventual use of both types of tests, the "adaptation" of currently available norm-referenced survey achievement tests is recommended.

BIBLIOGRAPHY


Arnold, R. "Reliability of test scores for the young 'bilingual' disadvantaged." The Reading Teacher, 1969, Vol. 22, No. 4.

Atilano, V. "Bilingual/Bicultural Education--An effective learning scheme for first grade Spanish-speaking, English-speaking, and American Indian children in New Mexico." ERIC ED 054 833, 1971.

Bernal, E. "Assessing Assessment Instruments: A Chicano Perspective." Unpublished manuscript, 1972.

Bordie, J. "Language Tests and Linguistically Different Learners: The Sad State of the Art." Elementary English, 1970, 1970, Vol. 47, No. 6, pg. 814-828.

Bortin, B. "Bilingual Program Evaluation Report, 1969-70." ERIC ED 043 708, 1970.

Buros, O. K. The Sixth Mental Measurements Yearbook. Highland Park, New Jersey: Gryphon Press, 1965.

_____. The Seventh Mental Measurements Yearbook. Highland Park, New Jersey: Gryphon Press, 1972.

California Department of Education. Personal correspondence to the author, May, 1972.

Cronbach, L. "Validity of Educational Measures," from Thorndike, R. (ed.), Educational Measurement. American Council on Education, 1970.

Davis, O. and Personke, C. "Effects of Administering the Metropolitan Tests in English and in Spanish to Spanish-Speaking School Entrants." Journal of Educational Measurement, 1968, Vol. 5, No. 3.

De Avila, E. "Multilingual Assessment: The Stockton Project." Paper presented at the National Conference on Bilingual Education. Austin, Texas; April, 1972.

Durost, W. "Directions for Administering: Metropolitan Achievement Tests, Primary II Battery." Yonkers on Hudson, New York: World Book Company, 1959.

Eagle, N. and Harris, A. "Interaction of Race and Test on Reading Performance Scores." Journal of Educational Measurement, 1969, Vol. 6, No. 3.

Ebel, R. "Criterion-referenced measurements: limitations." School Review, 1971, Vol. 79, 282-288.

Edmonston, L., Randall, R., and Oakland, T. "A Model for Estimating the Reliability and Validity of Criterion-Referenced Measures." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April, 1972.

Educational Testing Service (ETS). "A Description of the Anchor Test Study." U.S. Office of Education contract OEC-0-71-4758 (284), 1971.

Elementary and Secondary Education Act of 1965 (ESEA), Public Law 89-10, Section 116 : 22.

Frenner, J. "Criterion-Referenced Interpretations of Survey Achievement Tests." Test Development Memorandum TDM-72-1. Princeton, New Jersey: Educational Testing Service, 1972.

Glaser, R. "Instructional Technology and the Measurement of Learning Outcomes." American Psychologist, 1963, Vol. 18, 519-521.

Green, D. "Racial and Ethnic Bias in Test Construction." Adapted from final report of U.S. Office of Education contract OEC-9-70-0058 (057). Monterey, California: CTB/McGraw-Hill, 1972.

Harcourt, Brace, Jovanovich, Inc. "Report No. 7: Guidelines for Standardization Sampling." Metropolitan Achievement Tests Special Reports, 1970 Edition. June, 1971.

Hawkridge, D. et al. "A Study of Further Exemplary Programs of Education of Disadvantaged Children, Final Report." ERIC ED 036 668, 1969.

Hoepfner, R. CSE Elementary School Test Evaluations. Center for the Study of Evaluation, UCIA Graduate School of Education, Los Angeles, 1970.

Horn, T. "Three Methods of Developing Reading Readiness in Spanish-Speaking Children in the First Grade." The Reading Teacher, 1966, Vol. 20, No. 3.

Houston, S. "Cultural disadvantages: creativity, cooperation." Behavior Today, 1971, 2, (24), 3.

Hurt, M. and Mishra, S. "The Reliability and Validity of the Metropolitan Achievement Tests for Mexican-American Children." Educational and Psychological Measurement, 1970, Vol. 30.

Hutchison, J. "Reading tests and nonstandard language." The Reading Teacher, 1972, Vol. 25, No. 5.

Johnson, G. "Metropolitan Tests: Inappropriateness for ESEA Pupils." Integrated Education, Vol. 9, No. 6.

Livingston, S. "Criterion-Referenced Applications of Classical Test Theory." Journal of Educational Measurement, 1972, Vol. 9, No. 1.

Mitchell, B. Cited as personal correspondence by Davis and Personke. Journal of Educational Measurement, 1968, Vol. 5, No. 3, pg. 234.

Morris, J., Pestaner, M. and Nelson, A. "Mobility and Achievement." Journal of Experimental Education, 1967, Vol. 35, No. 4.

National Consortia for Bilingual Education (NCBE). Tests in Use in Title VII Bilingual Projects. Fort Worth, Texas; June, 1971.

Palomares, U. and Cummins, E. "Assessment of Rural Mexican-Americans in Preschool and Grades 1-6." ERIC ED 020 845, 1967.

Perrodin, A. and Snipes, W. "The Relationship of Mobility to Achievement in Reading, Arithmetic, and Language in Selected Georgia Elementary Schools." Journal of Educational Research, 1966, Vol. 59, No. 7.

Personke, C. and Davis, O. "Predictive Validity of English and Spanish Versions of a Readiness Test." The Elementary School Journal, 1969, Vol. 70, No. 2.

Pickering, C. "A Study of Intellectual Abilities of Culturally Disadvantaged Children as Predictors of Achievement in Reading, Mathematics, and Listening in Grade One." Doctor's thesis, Ohio University (Athens, Ohio), 1969 (DAI 31: 1085 A).

Popham, W. and Husek, T. "Implications of criterion-referenced measurement." Journal of Educational Measurement, 1969, Vol. 6, pg. 1-9.

Robison, A. "Relationship of measures of reading success of average, disadvantaged and advantaged Kindergarten children." The Reading Teacher, 1966, Vol. 20, No. 3.

Soloman, A. "The Effect of Answer Sheet Format on Test Performance by Culturally Disadvantaged Fourth Grade Elementary School Students." Journal of Educational Measurement, 1971, Vol. 8, No. 4.

Texas Education Agency (TEA). Annual Evaluation Report for Texas Child Migrant Program, 1969-70, Division of Assessment and Evaluation, 1970.

Wang, M. "Approaches to the validation of learning hierarchies." Western Regional Conference on Testing Problems (Proceedings) 1969. Princeton, New Jersey: Educational Testing Service, 14-38.

Williams, R. "Black Pride, Academic Relevance, and Individual Achievement." The Counseling Psychologist, 1970, Vol. 2, pg. 18-22.